A GMM Parts Based Face Representation for Improved Verification through Relevance Adaptation

Simon Lucey Tsuhan Chen

Advanced Multimedia Processing Laboratory, Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh PA 15213, USA
slucey@ieee.org, tsuhan@cmu.edu

Abstract

Motivated by the success of parts based representations in face detection we have attempted to address some of the problems associated with applying such a philosophy to the task of face verification. Hitherto, a major problem with this approach in face verification is the intrinsic lack of training observations, stemming from individual subjects, in order to estimate the required conditional distributions. The estimated distributions have to be generalized enough to encompass the differing permutations of a subject's face yet still be able to discriminate between subjects. In our work the well known Gaussian mixture model (GMM) framework is employed to model the conditional density function of the parts based representation of the face. We demonstrate that excellent performance can be obtained from our GMM based representation through the employment of adaptation theory, specifically relevance adaptation (RA). Our results are presented for the frontal images of the BANCA database.

1. Introduction

A problem of immense importance across the broad gamut of pattern recognition tasks at the moment is the ability to produce robust and well trained classifiers from sparse amounts of training observations. The task of face verification is a prime candidate for the development and analysis of this generic pattern recognition problem, as the nature of the task demands an ability to generalize for an infinite number of intra-class permutations from a finite and typically small facial image gallery set. A person's face is typically a varying object, more aptly described by a distribution rather than a static observation point. In this paper we present an approach that is able to estimate a distribution, from the subject's gallery image set, that is representative of most variations encountered in the probe set for that subject; whilst preserving the class distinction between subjects.

Distributions often occur as the consequence of collapsing some structural characteristic of an observation point. For example, if one considered an observation point as a sequence of face images taken over time, so that the observation point exists in an extremely high dimensional space[1], one could collapse the time structure to create a distribution of face images which is independent of time. However, this observation point would have to extend over a very long time sequence, with the sequence containing much variation, to capture all the facial permutations possible for a subject contained in such a high dimensional image. To overcome this problem one could also collapse some spatial structure in the image sequence. This would generate more observations and reduce the dimensionality further; resulting in a more generalized conditional distribution model. However, the collapse of some of the spatial structure in the image sequence may come at the cost of not being able to discriminate between face images stemming from different subjects.

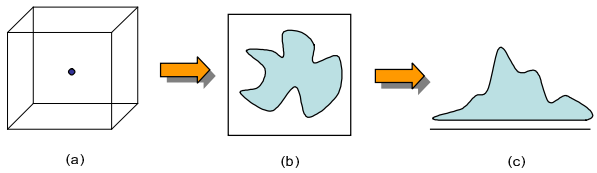

Figure 1: Graphical depiction on the effect of relaxing structural characteristics on an observation point. (a) Depicts the notion of a observation point existing in an extremely high dimensional space. (b) Depicts a collapse of "some" structure in the observation point, but still preserving some structure so it exists in a "moderately" sized dimensional space. (c) Depicts a "complete" collapse of structure in the observation point so that the distribution only exists in a "single"dimension.

Figure 1(a) depicts the extreme position of an observa-

[1]The image sequence in this sense exists in more than a 3D space. Each pixel and time stamp represents a dimension so that the entire sequence can be viewed as a single point in a extremely high dimensional space.

tion point existing in an extremely high dimensional space. In many classification tasks much of this representation's dimensionality may be redundant, detracting from our ability to match that observation point with similar observation points stemming from the same class. The distribution in Figure 1(c) depicts the other extreme, where "all" structure in the observation point has been collapsed onto a single dimension. The dense nature of the distribution will be more conducive to generalizing to other observation points from the same class. Unfortunately, this generality will most likely come at the cost of discriminating between observation points stemming from a different class. The situation in Figure 1(b) is the conceptual balance we are striving for in this paper, where sufficient distinction and generalization exists due to the collapse of "some" structure.

This concept has particular benefit in the task of face recognition. Collapsing some spatial structure in an image allows one to generate dense distribution's, parametric or nonparametric [1], describing gallery and probe image sets for a particular subject. The notion is especially powerful when the size of both the gallery and probe sets are meagre (i.e. a single image); as one can still compare distributions, borrowing on their natural ability to generalize, instead of individual points which can vary dramatically.

Previous work by Brunelli et. al [2] and Moghaddam et. al [3] cited good recognition performance by representing the face as a set of salient parts/regions (eg. eyes, nose, mouth). Images in the gallery set were used to create modular templates for comparison with salient regions of the probe images. Both Brunelli and Moghaddam noted superior performance by analyzing the image in a modular manner, rather than holistically as long as the salient regions had been localized to a satisfactory accuracy. Martínez [4] demonstrated a technique to model the uncertainty associated with the localization of these salient regions during the estimation of the modular templates. However, all these techniques essentially compare "points" (i.e. the distance from a probe's eye image to a eye template) not distributions. The work in this paper is motivated under the premise that the comparison of distributions have better generalization properties than the comparison of points; provided suitable class distinction is preserved.

The concept of reducing spatial structure in images, to gain a distribution, to aid in face classification tasks is not new. Much work has been done in the realm of face detection [5, 6], where benefit has been cited by viewing a face as being composed of both *parts* and *shape*. The *parts* are image patches containing information about the face in a local region. The *shape* component provides information describing where these patches are located globally within the face. By collapsing some of the *shape* structure of a face, accurate distributions can be estimated that generalize well to most permutations of faces whilst providing enough distinction between face and non-face regions in an image. The estimation of effective conditional face and non-face distributions requires the analysis of tens to hundreds of thousands of images.

Hitherto, a major problem in applying a similar philosophy to face verification is the typically small gallery set of images available for a subject. The lack of training observations drastically effects the ability to estimate conditional distributions that are generalized to differing permutations of a subject's face yet still contain enough complexity to discriminate between subjects. In this paper we present a technique, based on Bayesian learning, that is able to produce such distribution models. Employing a Gaussian mixture model (GMM) framework to model these distributions an adaptation technique, which we refer to as relevance adaptation (RA), is presented that can produce very complex but precise distributions for a subject from a small sized gallery set. These distributions provide a drastic improvement over techniques that do not employ adaptation. In this paper we have restricted our experiments to frontal images. Results are presented on the English portion of the BANCA database [7].

2. Model adaptation vs. estimation

Model adaptation [8], as the name suggests, implies that their is a pre-existing model whose parametric representation can be adapted from its current representation to describe a desired class of observations. The adaptation process is normally performed in such a way that the classification performance realized by the newly adapted model will be superior to the performance realized by the model if one was to perform estimation from scratch (i.e. train a model purely from a class-specific training set). Model adaptation is typically only of use, over estimation, when one has a limited amount of training observations for the class. In the presence of large amounts of training observations the need for adaptation generally disappears as all information about the model can be learnt from the abundant class-specific training observations.

The pre-existing model, required for adaptation, has usually been estimated from a training set, commonly referred to as a "development" set, that is several orders of magnitude larger than the actual class-specific training set. This development set usually subsumes or is at least representative of the *type* of classes (e.g. subject's faces) trying to be learnt from the training set, such that one can obtain statistics (i.e. a priori knowledge) about the nature of the class trying to be learnt. The term *adaptation* is employed instead of *estimation* as most of the prior density parameters are derived from parameters of the pre-existing model. In our work with frontal face verification the development set, used to estimate this initial model, stems from a reasonably

large population of subjects.

3. Estimating the distribution

Maximum a posteriori (MAP) estimation, or Bayesian estimation as it is commonly referred to [1], is a technique for estimating a distribution by employing a priori knowledge of how that model varies. Given that we have a set of training observations $\mathcal{S}_{trn}$ i.i.d. from an unknown distribution $f(\mathbf{o})$, but having an approximately known parametric form $\boldsymbol{\lambda}$, our task in MAP estimation is to find,

$$\boldsymbol{\lambda}_{MAP} = \arg\max_{\boldsymbol{\lambda}} f(\mathcal{S}_{trn}|\boldsymbol{\lambda})g(\boldsymbol{\lambda}) \quad (1)$$

where $g(\boldsymbol{\lambda})$ is the prior distribution governing how $\boldsymbol{\lambda}$ varies.

Often times in statistics, it is not easy to select an appropriate prior distribution. It is instead convenient to use an improper distribution (non-informative prior) that is represented by a nonnegative density function whose integral over the whole parameter space is infinite. We refer to this special case of MAP estimation as maximum likelihood (ML) estimation where all knowledge about $\boldsymbol{\lambda}$ stems from the observations. ML estimation is how we gain our initial model to be adapted.

3.1. Gaussian mixture models

The parametric form of $\boldsymbol{\lambda}$ will be a Gaussian mixture model (GMM). GMMs are employed in our work as they are able to provide a piece-wise parametric description of complex distributions using a number of relatively simple multivariate Gaussian distributions. GMMs are additionally attractive as they stem from the exponential family of distributions so that their well-known convexity property can be taken advantage of in estimation; a prime example being in use with the EM-algorithm [9].

A GMM models the probability distribution of a d dimensional statistical variable $\mathbf{o}$ as the sum of M multivariate Gaussian functions,

$$f(\mathbf{o}|\boldsymbol{\lambda}) = \sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \quad (2)$$

where $\mathcal{N}(\mathbf{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the evaluation of a normal distribution for observation $\mathbf{o}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The weighting of each mixture component is denoted by w_m and must sum to unity across all mixture components. The parameters of the model $\boldsymbol{\lambda} = \{w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}_{m=1}^{M}$ can be estimated using the Expectation Maximization (EM) algorithm [9] based on either a MAP or ML criterion. In the ML case, K-means clustering [1] was used to provide initial estimates of these parameters. In our work the covariance matrices in $\boldsymbol{\lambda}$ are assumed to be diagonal such that $\boldsymbol{\Sigma} = diag\{\boldsymbol{\sigma}^2\}$, as substantial benefit can be attained by reducing the number of parameters needing to be estimated.

3.2. Relevance adaptation

There are a variety of ways to gain a priori information about the distribution $g(\boldsymbol{\lambda})$. The employment of a *world*, or *universal background model* as it is sometimes referred to [10], has been shown empirically to greatly improve performance in GMM-based speaker verification. A world model is simply a single model trained from a large number of subject faces representative of the population of subject faces expected during verification, and usually has been estimated from a training set independent of the clients to be adapted. This world model is typically trained using the ML criterion (i.e. no informative prior).

Given a world model $\boldsymbol{\lambda}_w = \{w_{w_m}, \boldsymbol{\mu}_{w_m}, \boldsymbol{\Sigma}_{w_m}\}_{m=1}^{M}$ and training observations from a single client, $\mathbf{O} = [\mathbf{o}_1, \ldots, \mathbf{o}_R]$, using the iterative EM-algorithm one can obtain update equations that incorporate the a priori knowledge in the world model, to maximize the parametric representation of a GMM. We refer to the adaptation of the world model $\boldsymbol{\lambda}_w$ to produce a client model $\boldsymbol{\lambda}_c$ as relevance adaptation (RA). For RA this results in the following update equations[2],

$$w_{c_m} = \left[(1-\alpha_m^w)w_{w_m} + \alpha_m^w \frac{\sum_{r=1}^{R}\gamma_m(\mathbf{o}_r)}{\sum_{m=1}^{M}\sum_{r=1}^{R}\gamma_m(\mathbf{o}_r)}\right]\beta \quad (3)$$

$$\boldsymbol{\mu}_{c_m} = (1-\alpha_m^{\mu})\boldsymbol{\mu}_{w_m} + \alpha_m^{\mu}\frac{\sum_{r=1}^{R}\gamma_m(\mathbf{o}_r)\mathbf{o}_r}{\sum_{r=1}^{R}\gamma_m(\mathbf{o}_r)} \quad (4)$$

$$\boldsymbol{\sigma}_{c_m}^2 = (1-\alpha_m^{\sigma})\left(\boldsymbol{\sigma}_{w_m}^2 + \boldsymbol{\mu}_{w_m}^2\right) + \alpha_m^{\sigma}\frac{\sum_{r=1}^{R}\gamma_m(\mathbf{o}_r)\mathbf{o}_r}{\sum_{r=1}^{R}\gamma_m(\mathbf{o}_r)} - \boldsymbol{\mu}_{c_m}^2 \quad (5)$$

where $\gamma_m(\mathbf{o})$ is the occupation probability for mixture m and α_m^{ρ} is a weight used to tune the relative importance of the prior and is calculated via a relevance factor τ^{ρ} in,

$$\alpha_m^{\rho} = \frac{\sum_{r=1}^{R}\gamma_m(\mathbf{o}_r)}{\tau^{\rho} + \sum_{r=1}^{R}\gamma_m(\mathbf{o}_r)} \quad (6)$$

Different relevance factors can be estimated for the weights, means and variances respectively (i.e. $\rho \in \{w, \mu, \sigma\}$). In a similar fashion to work performed by Reynolds et. al [10] we have found effective performance can be attained by using a single relevance factor ($\tau = \tau^w = \tau^{\mu} = \tau^{\sigma}$). Empirically we found a relevance factor of $\tau = 16$ received good performance. The scale factor, β, in Equation 3 is computed to ensure that all the adapted mixture component weights sum to unity. Finally, it must be noted that the adaptation framework presented in this

[2] A derivation of Equations 4 and 5 was developed by Gauvain and Lee [8]. The weight update in Equation 3 was found experimentally to perform better than Gauvain and Lee's original.

paper differs marginally to that presented by Reynolds as the updates, performed in Equations 3-5, are done at each iteration of the EM algorithm. This was done as the resultant models were found to be more stable and effective than applying the updates after the iterative process.

3.3. Evaluation

When evaluating a sequence of observations, from a claimant, we obtain the average log-likelihood,

$$\mathcal{L}(\mathbf{O}|\boldsymbol{\lambda}_c) = \frac{1}{R}\sum_{r=1}^{R}\log f(\mathbf{o}_r|\boldsymbol{\lambda}_c) \quad (7)$$

Given the average log-likelihood, for the client and world models, one can then calculate the log-likelihood ratio,

$$\Lambda(\mathbf{O}) = \mathcal{L}(\mathbf{O}|\boldsymbol{\lambda}_c) - \mathcal{L}(\mathbf{O}|\boldsymbol{\lambda}_w) \quad (8)$$

Verification is performed by accepting a claimant when $\Lambda(\mathbf{O}) \geq Th$ and rejecting him/her when $\Lambda(\mathbf{O}) < Th$, where Th is a given threshold. Verification performance is evaluated using two measures; being false rejection rate (FRR), where a true client is rejected against their own claim, and false acceptance rate (FAR), where an impostor is accepted as the falsely claimed client. The FAR and FRR measures increase or decrease in contrast to each other based on the threshold Th. A simple measure for overall performance of a verification system is found by determining the equal error rate (EER) for the system, where FAR = FRR.

4. *Parts* and feature representations

An initial investigation into what features are most effective for the *parts* representation of frontal face image was conducted by Sanderson et. al [11] for the task of face verification. Sanderson's work is pertinent to our work as it was one of the first investigations for *parts* based face verification using GMMs; albeit using a ML criterion. In this work a modified form of the 2D discrete cosine transform (2D-DCT) was recommended, in comparison to other representations like 2D-Gabor features, as an ideal way to gain a compact *parts* representation that provided good distinction between the faces of subjects and fast feature computation. A depiction of the feature extraction process can be seen in Figure 2.

The experiments conducted in this paper were performed on cropped faces geometrically normalized for rotation and scale so as to form an 114×91 array of pixels. These images were also statistically normalized to have a unit variance and zero mean. The face images were then decomposed into 16×16 pixel image patches with an overlap between horizontal and vertical adjacent patches of 75%. The overlap between patches aids verification from two perspectives.

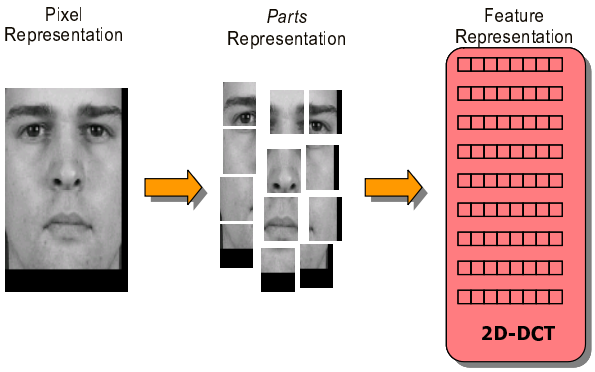

Figure 2: Graphical depiction of the parts and feature representations of a face. Note: even though overlapping blocks are not depicted in practice the overlapping of blocks leads to greater performance.

First, the overlap reduces the spatial area used to derive one feature vector and adds some redundancy between patches (i.e. no single patch contains all the information about a local region of the face). Second, as the overlap is increased it also increases the number of image patches (i.e. observations) exponentially.

Once the image patches are acquired they then have an 2D-DCT applied to compact the $16 \times 16 = 256$ element patch into a feature vector $\mathbf{o}$ of suitable dimensionality to model a generalized but distinct distribution of that subject's face. The first 64 energy preserving 2D-DCT coefficients are extracted, according to a zig-zag pattern [11], with the first coefficient being dropped as it is represents the mean of the patch. We found empirically that the removal of the first coefficient improved verification performance. This results in a $d = 63$ observation feature vector $\mathbf{o}$ for used in adaptation.

5. BANCA database

The English portion of the BANCA database was employed for these experiments containing 52 subjects; evenly divided into two sets $[g1, g2]$ of 26 as per the BANCA protocol [7]. Inside those sets there are an equal number of sexes (i.e. male=13, female=13). The $g1$ and $g2$ sets are used for the development and evaluation sets in our experiments. The development set is used to obtain any data-dependent aspects of the verification system (e.g. world model etc.). The evaluation set is where the performance rates for the verification system are obtained.

If the $g1$ set is used as the development set then the $g2$ set is used for the evaluation set; and vice versa. This is done to avoid any methodological flaw, as it is essential that the development set is composed of a distinct subject population

as the one of the evaluation set. We will report results in this paper using both the $g1$ and $g2$ sets so as to gain a gauge for the statistical significance of our results. Several protocols [7] have been devised for the BANCA database. For the experiments in this correspondence we have employed the "matched conditions" (MC) protocol where images in the gallery and probe sets stem from the same camera under the same conditions. There are a total of 4 sessions used in the protocol with the first session being used as the gallery with the remaining 3 session being used for the probe. Each session consisted of a sequence of 5 images per subject, taken as the subject is speaking. In the BANCA database each subject has his/her session recorded with a *client access* utterance and a *imposter attack* utterance. The client access utterance sessions 2,3 and 4 were used only for client verification with the imposter attack utterance sessions being taken from all 4 sessions.

6. Full shape collapse

A question now presents itself on how much structure should we collapse in a face image? Given that we have a *parts* and *shape* representation of the face, we can choose to collapse all the shape structure in the representation; which we shall refer to as full shape collapse (FSC). In our work we investigated the performance of a FSC-GMM based face verification scheme employing RA on all of the parameters $(w, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ of each mixture component (i.e. applying Equations 3-5). We also investigated applying the RA scheme on the means $(\boldsymbol{\mu})$ only of each mixture component (i.e. applying Equation 4). This was motivated by the benefit seen in previous GMM based work [10] of reducing the number of parameters needing to be estimated. Both these schemes were evaluated using $\tau = 0$ (i.e. ML estimation) and $\tau = 16$ (i.e. MAP estimation) for the $g1$ and $g2$ BANCA evaluation sets. In Figure 3 verification results are presented as a function of the number of mixture components (M).

A log-scale was employed in Figure 3 to evaluate verification performance from a simple (e.g. 4) to a very complex (e.g. 2048) value of M. It is obvious for the $\{\tau = 0 - (w, \boldsymbol{\mu}, \boldsymbol{\Sigma})\}$ strategy that there is a drastic deterioration in performance from increasing the complexity of the GMMs. The poor performance for high values of M can be attributed to the problem we stated at the beginning of this paper concerning the estimation of generalized but discriminative distributions from a small sized gallery set.

Good performance was attained for the $\{\tau = 0 - (\boldsymbol{\mu})\}$ strategy for low values of $M < 512$. Superior performance was achieved for the $\{\tau = 16 - (\boldsymbol{\mu})\}$ and $\{\tau = 16 - (w, \boldsymbol{\mu}, \boldsymbol{\Sigma})\}$ strategies employing increasingly complex GMMs. A monotonic improvement in performance was seen as a function of M for the two strategies. Minimal

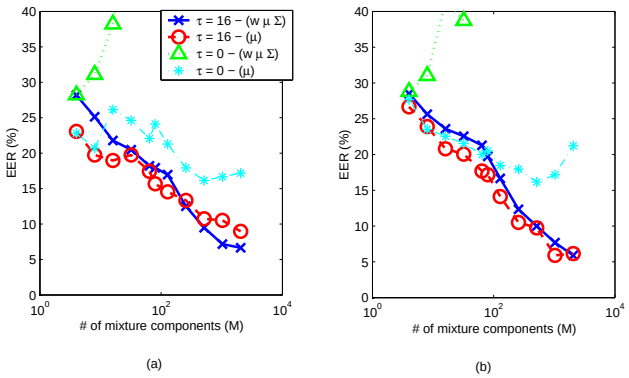

Figure 3: Comparison of various adaptation schemes across the (a) $g1$ and (b) $g2$ BANCA evaluation sets.

performance benefit was attained by setting $M > 2048$. A clear benefit in performance was witnessed for the two strategies employing non-zero values of τ; that is using a MAP rather than ML criterion for estimation. For both the $g1$ and $g2$ BANCA sets slightly better performance was achieved using a $\{\tau = 16 - (w, \boldsymbol{\mu}, \boldsymbol{\Sigma})\}$ for $M = 2048$.

7. Partial shape collapse

An obvious question stemming from the previous section is whether there is any benefit in keeping "some" shape structure in the face representation? Even though we can see that good performance can be attained by collapsing the shape structure completely in a face, "some" form of simplified shape structure may be beneficial to verification performance. For example, there may be benefit in enforcing that fiducial regions of face images be compared against each other (e.g. eye region against eye region, mouth region against mouth region, etc.). Motivated by this concept we have attempted to place some labels on regions of the face. We refer to this label based representation as partial shape collapse (PSA). Figure 4 contains some examples of how faces were labelled.

These labels specifically refer to the facial objects $q \in \{$ brow, left eye, right eye, bridge of nose, nose, left cheek, right cheek, mouth$\}$. A region was defined for each label using a single Gaussian distribution. The mean of the Gaussian, based on hand-labelled coordinates, was centered within the labelled object. The covariance matrices of each Gaussian were heuristically chosen to encompass the approximate area of the labelled object. Patches were marked as belonging to an object label if the patch was located within the 90% ellipsoid boundary of the Gaussian.

Using these labels we then attempted to estimate separate conditional distribution models $\boldsymbol{\lambda}_{c,q}$ based on the identity c and *shape* label q of the *parts*. During verifica-

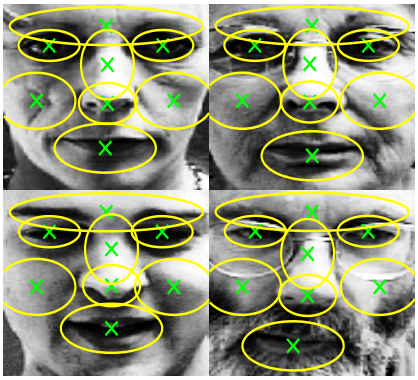

Figure 4: Examples of how frontal faces were labelled using $8\times$Gaussian distributions.

tion this results in a conditional log-likelihood ratio $\Lambda_q(\mathbf{O})$ for each label on the face. In our work, after the estimation/adaptation of $\boldsymbol{\lambda}_{c,q}$, a log-likelihood ratio for a claimant's entire face is calculated as,

$$\Lambda(\mathbf{O}) = \sum_q \Lambda_q(\mathbf{O}) \tag{9}$$

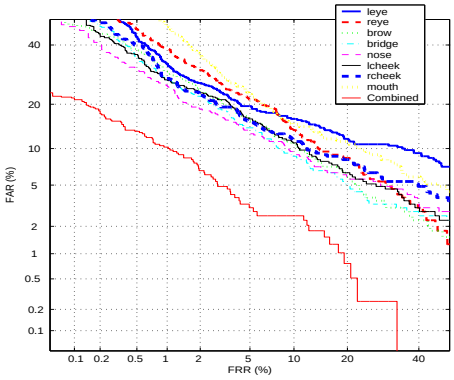

Figure 5: DET curve depicting the distinction between different labels on the face, using the $g1$ set, using $M = 256$ mixture components for each facial region GMM.

In Figure 5 we see the detection error tradeoff (DET) curve of the distinction provided by each label for our best performing scheme using $M_q = 256$ mixture components for each facial region's GMM. Interestingly one can see relatively homogeneous regions like the cheeks, bridge and brow providing better verification performance than inhomogeneous regions like the mouth and to a lesser extent the eyes. One can also see the clear benefit in combining the scores with verification performance outperforming all the regions individually.

For simplicity, in our experiments, we set M_q to be the same across all the facial regions. Future work may find benefit in applying differing values for M_q to each region. Like the previous section we performed a log-scale exhaustive search on what values of M_q perform best for PSC-GMMs using a relevance factor $\tau = 16$. The results of this search can be seen in Figure 6 relative to the FSC-GMM technique. The value M used in Figure 6, for the PSC-GMM technique, is the total number of mixtures used (i.e. $M = 8 \times M_q$). In our initial experiments, across $g1$ and $g2$ sets, we found no real benefit in evaluating $M_q > 512$. Similar results were received to those seen with FSC-GMMs with marginally better performance being received for $\{\tau = 16-(w, \boldsymbol{\mu}, \boldsymbol{\Sigma})\}$ over the $\{\tau = 16-(\boldsymbol{\mu})\}$ strategy. For an undetermined reason the margin of performance improvement for the PSC-GMM over the FSC-GMM was greater for set $g1$ than $g2$.

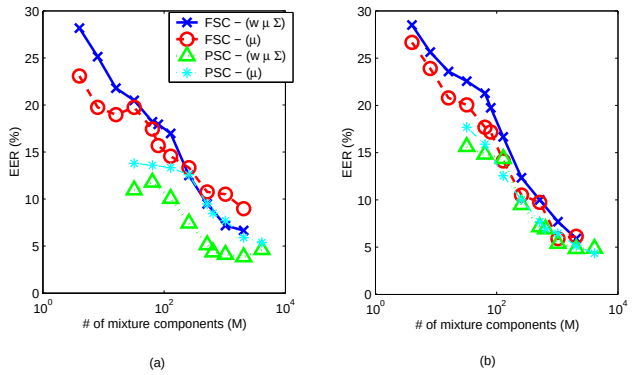

Figure 6: Comparison of verification performance for FSC-GMM and PSC-GMM techniques for the (a) $g1$ and (b) $g2$ BANCA evaluation sets.

8. Baseline comparison

Although the explicit purpose of this paper is to elucidate, through adaptation theory, upon a different perspective to face verification, it is also important to provide a baseline comparison to the work we have presented. We conducted an experiment to compare our FSC and PSC-GMM algorithms with two others. In particular we compared our algorithm with Eigenfaces [12] and Fisherfaces [13]. Eigenfaces and Fisherfaces are the defacto baseline standard by which face recognition algorithms are typically compared.

One can see both the FSC and PSC-GMM algorithms, across both the $g1$ and $g2$ BANCA sets, provide, by a large margin, superior performance to the baseline Eigenface and Fisherface algorithms. As expected the PSC-GMM algorithm outperforms the FSC-GMM algorithm, but only by a marginal amount. This very interesting result indicates that

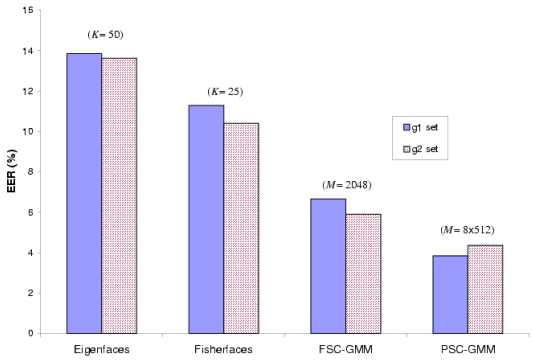

Figure 7: Comparison of FSC and PSC-GMM algorithms, using M mixture components, against baseline Eigenface and Fisherface algorithms, using K eigenvectors.

one may in some instances be able to obtain good verification performance without having to accurately locate local facial features (e.g. eyes, mouth, etc.).

9. Summary and Conclusions

In this paper we have presented a new perspective on the task of face verification that demonstrates good performance through the collapse of structure via a *parts* and *shape* representation of a face image. RA, through a MAP criterion, has shown to be of benefit in estimating complex GMMs for use in face verification, circumventing many of the problems associated with using a ML criterion. One of the major contributions of our work is to demonstrate how RA can be employed to train very complex (e.g. M=2048) GMMs for use in verification. Additionally, results indicate that simplifications to what parameters are adapted for client GMMs (i.e. adapting the means only) has minimal effect on verification performance for both FSC and PSC-GMMs. These simplifications may aid in the formulation of future adaptation work that tries to take advantage of any dependencies that exist between mixture components.

The ability to estimate complex distributions, through RA, has allowed us to explore a face verification paradigm using a *parts* philosophy for representing the face. The results received for the FSC-GMM algorithm are particularly encouraging as they indicate that one may not need to locate local facial features to receive good verification performance when performing verification by *parts*.

Acknowledgments

This research is supported in part by the Technology Support Working Group (TSWG).

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY, USA: John Wiley and Sons, Inc., 2nd ed., 2001.

[2] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE Trans. PAMI*, vol. 15, no. 10, pp. 1042–1052, 1993.

[3] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object recognition," *IEEE Trans. PAMI*, vol. 19, no. 7, pp. 696–710, 1997.

[4] A. M. Martínez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Trans. PAMI*, vol. 24, no. 6, pp. 748–763, 2002.

[5] M. Weber, M. Welling, and P. Perona, "Towards automatic discovery of object categories," in *CVPR*, pp. 101–108, June 2000.

[6] H. Schneiderman and T. Kanade, "A histogram-based method for detection of faces and cars," in *CVPR*, pp. 504–507, September 2000.

[7] E. Bailly-Bailliere, S. Bengio, K. Messer, and et. al, "The BANCA database and evaluation protocol," in *AVBPA*, 2003.

[8] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. SAP*, vol. 2, no. 2, pp. 291–298, 1994.

[9] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[11] C. Sanderson and K. Paliwal, "Fast features for face authentication under illumination direction changes," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2409–2419, 2003.

[12] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, 1991.

[13] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. PAMI*, vol. 19, no. 7, pp. 711–720, 1997.